



NEWS RELEASE

QumulusAI Signs More Than \$124 Million in AI Inference Infrastructure Agreements

2026-06-11

Workload-optimized Nvidia Blackwell deployments designed to reduce AI inference costs by approximately 20% compared with standard reference architectures

ATLANTA, GA / [ACCESS Newswire](#) / June 11, 2026 / [QumulusAI](#), a vertically integrated AI cloud infrastructure company, today announced it has secured more than \$124 million in customer subscriptions for 3-year terms with [Hyperbolic](#) and another leading AI inference platform. These agreements support deployments totaling 1,280 NVIDIA Blackwell GPUs and include nearly \$21.9 million in combined upfront customer commitments.

The customers operate some of the industry's largest inference platforms for open-source AI models, powering deep-research agents, automated coding systems, and other asynchronous AI applications that require high-throughput, low-latency and cost-efficient compute infrastructure. The agreements establish long-term recurring revenue for QumulusAI and validate growing demand for infrastructure purpose-built on AI inference workloads for these applications.

Under the GPU-as-a-service agreements, QumulusAI will provision 160 Lenovo and Supermicro bare-metal servers equipped with NVIDIA B300 and B200 Blackwell GPUs, respectively. [Cisco Nexus One](#) will power the cluster fabric for both deployments, delivering secure, high-performance AI networking. Rather than deploying off-the-shelf reference builds, QumulusAI's deployment is designed to reduce AI inference costs by approximately 20% compared to standard configurations by rightsizing CPU core counts, system memory, and local storage to the exact demands of large-scale open-source inference.

"AI infrastructure can no longer be built using one-size-fits-all designs," said Mike Maniscalco, CEO of QumulusAI. "Inference workloads have very different performance and economic requirements than model training environments. By tuning infrastructure to the workload itself, we can improve utilization, reduce costs, and accelerate deployment timelines for customers operating at production scale."

"As AI adoption expands, organizations need access to infrastructure that can be deployed quickly, scaled efficiently, and aligned to the economics of production AI," added Maniscalco. "These agreements demonstrate the value of a more flexible approach to AI infrastructure."

One of the agreements is with [Hyperbolic](#), an AI cloud platform focused on providing scalable GPU compute infrastructure for AI startups, research teams, and enterprises. Hyperbolic gives AI builders flexible access to reliable, cost-efficient compute for training, fine-tuning, and inference workloads, helping teams move faster from experimentation to production.

"AI teams need infrastructure that supports every stage of the AI lifecycle, from training and fine-tuning to production inference," said Jasper Zhang, CEO of Hyperbolic. "QumulusAI's workload-optimized infrastructure gives us the performance, efficiency, and scalability we need as we continue expanding reliable GPU compute for customers building AI at scale."

About QumulusAI

QumulusAI is a distributed AI cloud platform that delivers accelerated access to high-performance GPU compute. Through an inference-first, demand-led deployment model across a network of data center sites, QumulusAI brings compute closer to customer demand, helping AI teams and enterprises scale production AI workloads with speed, flexibility and control. By combining rapid deployment with flexible private cloud infrastructure, QumulusAI gives customers a faster, more adaptable path beyond the capacity constraints of traditional centralized and hyperscale cloud models. Learn more at [QumulusAI.com](https://qumulusai.com).

Investor Contact

investors@qumulusai.com

Media Contact

media@qumulusai.com

Follow QumulusAI on social media: <https://www.linkedin.com/company/qumulusai>

Disclaimer: This press release contains certain "forward-looking statements" that are based on current expectations, forecasts and assumptions that involve risks and uncertainties, and on information available to QumulusAI as of the date hereof. QumulusAI's actual results could differ materially from those stated or implied herein, due to risks and uncertainties associated with its business. Forward-looking statements include statements regarding QumulusAI's expectations, beliefs, intentions or strategies regarding the future, and can be identified by forward-looking words such as "anticipate," "believe," "could," "continue," "estimate," "expect," "intend," "may," "should," "will" and "would" or words of similar import. QumulusAI expressly disclaims any obligation or undertaking to disseminate any updates or revisions to any forward-looking statement contained in this press release to reflect any change in QumulusAI's expectations with regard thereto or any change in events, conditions or circumstances on which any such statement is based in respect of its business, partnerships or otherwise.

SOURCE: QumulusAI

View the original [press release](#) on ACCESS Newswire